Is Parameter Learning via Weighted Model Integration Tractable?

Abstract

Weighted Model Integration (WMI) is a recent and general formalism for reasoning over hybrid continuous/discrete probabilistic models with logical and algebraic constraints. While many works have focused on inference in WMI models, the challenges of learning them from data have received much less attention. Our contribution is twofold. First, we provide novel theoretical insights on the problem of estimating the parameters of these models from data in a tractable way, generalizing previous results on maximum-likelihood estimation (MLE) to the broader family of log-linear WMI models. Second, we show how our results on WMI can characterize the tractability of inference and MLE for another widely used class of probabilistic models, Hinge Loss Markov Random Fields (HL-MRFs). Specifically, we bridge these two areas of research by reducing marginal inference in HL-MRFs to WMI inference, and thus we open up new interesting applications for both model classes.

1 INTRODUCTION

Solving many complex AI tasks require both logical and numerical reasoning in uncertain environments. Probabilistic inference in hybrid continuous/discrete settings pose significant challenges, in particular when logical and algebraic constraints are considered, such as laws of physics or safety requirements. In many discrete probabilistic models, reducing marginal inference to Weighted Model Counting (WMC) [Sang et al., 2005] is a state-of-the-art technique, thanks to its support for arbitrary propositional logic constraint. Recently, Weighted Model Integration (WMI) [Belle et al., 2015] emerged as a generalization of WMC to hybrid domains. By extending propositional logic with algebraic relations over continuous variables and generalizing the weight function to piecewise densities, WMI unifies marginal inference in a large class of probabilistic models under the same formalism.

WMI-based inference has been receiving increasing interest in recent years, with advancements both on practical algorithms [Morettin et al., 2021] and on theoretical aspects [Zeng et al., 2020a]. While its theoretical groundwork was initially developed to address inference and parameter learning in hybrid Markov Random Fields, over the last five years it has been shown that a much larger class of problems can be reduced to WMI, such as marginal inference in Mixed Sum-Product Networks [Morettin et al., 2020], or verification of fairness properties in probabilistic programs [Albarghouthi et al., 2017]. Yet, many questions related to learning this class of distributions from data are still open.

In this paper, we investigate weight learning for WMI models in a principled way. First, we show under which conditions weight learning can be done in a tractable way and by doing this we greatly extend the class of WMI models amenable to efficient learning. Additionally, we show how a popular class of probabilistic models, namely Hinge Loss Markov Random Fields (HL-MRFs) [Bach et al., 2017], fall into this class by reducing marginal inference in HL-MRFs to WMI. Leveraging recent results on the tractability of WMI-based inference, we then characterize a subclass of HL-MRFs that admit tractable maximum likelihood estimation of their parameters and marginal inference.

2 BACKGROUND

Notation. Uppercase letters denote random variables (X, B) and lowercase letters denote their assignments (x, b). We use bold for sets of variables $(\boldsymbol{X}, \mathbf{B})$, and their joint assignments $(\boldsymbol{x}, \boldsymbol{b})$. We use capital Greek letters for logical formulas (Γ, Δ). Literals are atomic formulas or their negation, and are denoted using either ℓ or lowercase Greek letters (γ, δ). We let $\boldsymbol{x} \models \Delta$ denote the satisfaction of a

Submitted to the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021). **To be used for reviewing only**.

formula Δ by an assignment $\boldsymbol{x}$. Its corresponding indicator function is $[\![\boldsymbol{x} \models \Delta]\!]$.

A state-of-the-art approach for answering probabilistic queries in many discrete models reduces the problem to *weighted model counting* (WMC), i.e. the task of computing the weighted sum of the models (solutions) of a propositional formula Δ:

$$\mathsf{WMC}(\Delta, \mathcal{W}) = \sum_{\mu \models \Delta} \mathcal{W}(\mu) \tag{1}$$

Typically, it is assumed that the weight of a model factorizes as the product of non-negative constant weights associated with each literal (an atom or its negation) in the solution, i.e. $\mathcal{W}(\mu) = \prod_{\ell \in \mathcal{L}} w_\ell^{[\![\mu \models \ell]\!]}$, where $[\![.]\!] : \mathbb{B} \to \{0, 1\}$ is the indicator function, $\mathcal{L} = Atoms(\Delta) \cup \{\neg A \mid A \in Atoms(\Delta)\}$ denotes the set of all the literals and $w_\ell \in \mathbb{R}^+$ is the positive weight of a literal ℓ (unweighted literals are assumed to have weight 1).

Example 1 Consider the formula $\Delta = (A \to B)$ and weights $w_A = 2, w_{\neg A} = 3, w_B = 5$, then:

$$\mathsf{WMC}(\Delta, \mathcal{W}) = \overbrace{2 \cdot 5}^{\mathcal{W}(A \wedge B)} + \overbrace{3 \cdot 5}^{\mathcal{W}(\neg A \wedge B)} + \overbrace{3 \cdot 1}^{\mathcal{W}(\neg A \wedge \neg B)} = 28.$$

The pair $\langle \Delta, \mathcal{W} \rangle$ encodes an unnormalized joint distributions over the propositional variables in Δ, thus the normalized probability of a formula Δ_Q given evidence expressed as another formula Δ_E is computed as:

$$Pr_{\langle \Delta, \mathcal{W} \rangle}(\Delta_Q | \Delta_E) = \frac{\mathsf{WMC}(\Delta \wedge \Delta_Q \wedge \Delta_E, \mathcal{W})}{\mathsf{WMC}(\Delta \wedge \Delta_E, \mathcal{W})}. \tag{2}$$

Algorithmic advances in both exact and approximate WMC motivated its generalization to hybrid continuous/discrete settings [Belle et al., 2015].

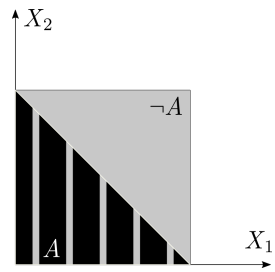
Given a set of continuous random variables $\mathbf{X}$ and a set of Boolean variables $\mathbf{A}$, the support of the joint probability $Pr(\mathbf{X}, \mathbf{A})$ is encoded by means of a *satisfiability modulo theories* (SMT) [Barrett et al., 2010] formula, i.e. a (typically quantifier-free) expression containing both propositional and *theory* atoms connected with the usual logical connectives. Specifically, the theory atoms encode algebraic constraints over $\mathbf{X}$, often restricted to the theory of *linear algebra over reals* ($\mathcal{LRA}$), where atoms have form $(\mathbf{c}^T \mathbf{X} \leq b)$.

Example 2 The following SMT-$\mathcal{LRA}$ formula:

$$\Delta = (0 \leq X_1) \wedge (X_1 \leq 1) \wedge (0 < X_2) \wedge (X_2 \leq 1) \\ \wedge (A \to (X_1 + X_2 \leq 1))$$

has 3 satisfying truth assignments ($X_1, X_2 \in [0, 1]$ are always true):

$$\{\overbrace{A \wedge (X_1 + X_2 \leq 1) \wedge ...}^{\mu_1}, \overbrace{\neg A \wedge (X_1 + X_2 \leq 1) \wedge ...}^{\mu_2}, \\ \overbrace{\neg A \wedge \neg(X_1 + X_2 \leq 1) \wedge ...}^{\mu_3}\}$$



for a total (unweighted) volume of $\frac{3}{2}$.

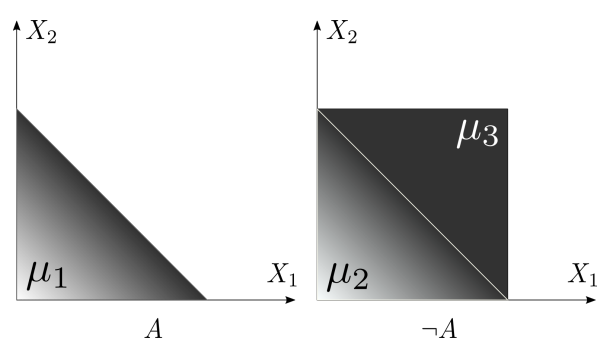
In addition to the usual constant weights associated with propositional literals, algebraic literals ℓ on variables $\mathbf{X}_\ell \subseteq \mathbf{X}$ can be mapped to functions over $\mathbf{X}_\ell$. Thus, the weight function $\mathcal{W}(\mu, \boldsymbol{x}) = \prod_{\ell \in \mathcal{L}} w_\ell(\boldsymbol{x}_\ell)^{[\![\mu, \boldsymbol{x} \models \ell]\!]}$ is a piecewise function, associating an unnormalized density to each solution $\mu \models \Delta$. In this context $\mathcal{L}$ denotes the set of weighted literals, which doesn't necessarily have to appear in Δ. Intuitively, a truth assignment to the literals $\mathcal{L}$ univocally defines a region of the piecewise distribution. Similarly to WMC, the weighted model integral is computed by summing over the solutions of the SMT formula and, additionally, by integrating the density function in the continuous domain:

$$\mathsf{WMI}(\Delta, \mathcal{W}) = \sum_{\mu \models \Delta} \int_{\boldsymbol{x} \models \mu} \mathcal{W}(\mu, \boldsymbol{x}) \, d\boldsymbol{x} \tag{3}$$

Piecewise polynomials, which admit closed-form solutions when integrated over arbitrary polytopes, are the most investigated class of joint density in the WMI literature.

Example 3 Consider the formula in Ex.2 and a single weighted literal $\ell = (X_1 + X_2 \leq 1)$ with $w_\ell(X_1, X_2) = X_1 + X_2$, i.e. $\mathcal{W}(X_1, X_2) = (X_1 + X_2)^{[\![X_1 + X_2 \leq 1]\!]}$. Then, the weighted model integral is:

$$\begin{aligned} \mathsf{WMI}(\Delta, \mathcal{W}) = &\int_{\mu_1} X_1 + X_2 \, dX_1 \, dX_2 \\ &+ \int_{\mu_2} X_1 + X_2 \, dX_1 \, dX_2 \\ &+ \int_{\mu_3} 1 \, dX_1 \, dX_2 = 2 \cdot \frac{1}{3} + \frac{1}{2} \end{aligned}$$



W.l.o.g., for notational convenience from here on we will focus on WMI problems on continuous variables only. This is possible since any WMI problem on continuous and Boolean variables can be reduced in polytime to a new WMI problem on continuous variables only, without changing its inference complexity class, by properly introducing

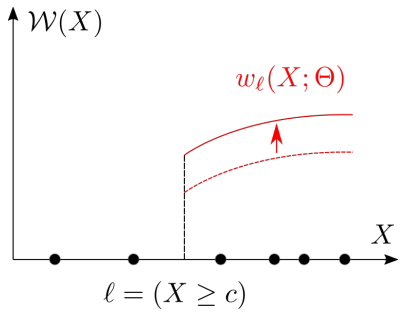


Figure 1: Graphical representation of the parameter estimation task considered in this work.

auxiliary variables to account for Boolean variables [Zeng and Van den Broeck, 2019]. We remark that the results on parameter learning reported in the following sections hold for hybrid distributions too. In fact, weighted propositional variables would contribute with piecewise constant factors only, a subcase of our generalized theory that was previously investigated by Belle et al. [2015]. We define the normalized density of any data point $\boldsymbol{x}$ as:

$$\mathsf{p}(\boldsymbol{x}) = \frac{\prod_{\ell \in \mathcal{L}} w_\ell(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell]\!]}}{\mathsf{WMI}(\Delta, \mathcal{W})}, \quad \boldsymbol{x} \models \Delta \tag{4}$$

Characterizing the dependency structure among the variables in a model is crucial for characterizing the tractability of probabilistic inference. This structure can be captured with the notion of *primal graph*.

Definition 2.1. *(Primal Graph) Given a model $\langle \Delta, \mathcal{W} \rangle$, its primal graph $\mathcal{G}_{\langle \Delta, \mathcal{W} \rangle} = (\mathcal{V}, \mathcal{E})$ is the undirected graph whose vertex set $\mathcal{V}$ is the index set of variables in formula Δ and $\mathcal{L}$, and whose edge set $\mathcal{E}$ has edge $i - j$ iff variable X_i and variable X_j appear together in one clause $\Gamma \in \Delta$ or in one weighted literal $\ell \in \mathcal{L}$.*

3 MAXIMUM LIKELIHOOD ESTIMATION OF WEIGHTED MODEL INTEGRATION PARAMETERS

We now consider the problem of estimating the parameters of a structured, piecewise distribution from data. Specifically, we are concerned with learning the parameters in the literal weights w_ℓ for $\ell \in \mathcal{L}$ (Fig. 1), assuming that the SMT formula Δ and the set $\mathcal{L}$ is fixed and that a dataset consisting of fully observed samples is provided.

We adopt the canonical parameter estimation approach, maximum likelihood estimation (MLE), that is,

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}) \tag{5}$$

where $L(\boldsymbol{\Theta})$ is the log-likelihood of the dataset, i.e.,

$$L(\boldsymbol{\Theta}) = \log \prod_{\boldsymbol{x} \in \mathcal{D}} \mathsf{p}(\boldsymbol{x}; \boldsymbol{\Theta}).$$

Plugging in the point-wise density p in Eq. 4 we get:

$$L(\boldsymbol{\Theta}) = \sum_{\boldsymbol{x} \in \mathcal{D}} \sum_{\ell \in \mathcal{L}} \log w_\ell(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell]\!]} - |\mathcal{D}| \cdot \log \mathsf{WMI}(\Delta, \mathcal{W}; \boldsymbol{\Theta})). \tag{6}$$

In general, the optimization problem in Equation 5 does not have a closed form solution. For this reason, we consider optimizing our objective via iterative methods, such as gradient ascent. We then need to analyze the mathematical properties of the log-likelihood $\frac{\partial L(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$ for characterizing convergence and tractability of the optimization problem.

Given the SMT formula Δ and the set of weighted literals $\mathcal{L}$, we explore whether the log-likelihood objective is concave, which guarantees convergence to the global optimum when iterative methods such as gradient ascent are used [Nesterov and Nemirovskii, 1994].

As reported in the original WMI paper [Belle et al., 2015], MLE of the parameters in weight functions can be solved optimally using standard convex optimization tools for constant weight functions, that is, when the function $w_\ell(\boldsymbol{x})$ is constant for any $\ell \in \mathcal{L}$. We generalize this result to weight functions that are *log-linear with respect to their parameters*. Specifically, the weight functions take their form as

$$w_\ell(\boldsymbol{x}_\ell; \boldsymbol{\Theta}_\ell) = \exp\{\boldsymbol{\Theta}_\ell^\top \boldsymbol{f}_\ell(\boldsymbol{x}_\ell)\}$$

where $\boldsymbol{f}_\ell(\cdot) = (f_\ell^1(\cdot), \cdots, f_\ell^K(\cdot))^\top$ is a vector of features defined over variable subset $\boldsymbol{X}_\ell$. Log-linear functions are an expressive weight family that generalizes many functions adopted in previous works. Some example choices for feature functions $\boldsymbol{f}(\cdot)$ could be 1) constant functions, in which case the WMI problem has constant weights, 2) monomials, in which case the WMI problem has exponentiated polynomial weights, 3) log-polynomials, in which case the WMI problem has polynomial weights.

Notice that the first term in the log-likelihood in Equation 6 is a linear combination of parameters, i.e.,

$$\sum_{\boldsymbol{x} \in \mathcal{D}} \sum_{\ell \in \mathcal{L}} \log w_\ell(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell]\!]} = \sum_{\boldsymbol{x} \in \mathcal{D}} \sum_{\ell \in \mathcal{L}} [\![\boldsymbol{x} \models \ell]\!] \cdot \Theta_\ell^\top f_\ell(\boldsymbol{x}_\ell).$$

We thus investigate the convexity of the log-partition function term $Z(\boldsymbol{\Theta}) = \mathsf{WMI}(\Delta, \mathcal{W}; \boldsymbol{\Theta})$.

Proposition 3.1. *The log-partition function $\log Z(\boldsymbol{\Theta})$ is convex if the weight functions are log-linear in their parameters.*

Sketch of Proof Denote the i-th parameter in weight function w_ℓ as Θ_i^ℓ. The first-order derivative of the log-partition function $\log Z(\boldsymbol{\Theta})$ with respect to a parameter Θ_i^ℓ is as follows.

$$\frac{\partial}{\partial \Theta_i^\ell} \log Z(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{x} \sim \mathsf{p}}[\frac{\partial}{\partial \Theta_i^\ell} \log w_\ell(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell]\!]}]$$

For brevity, denote $\frac{\partial}{\partial \Theta_i^\ell} \log w_\ell(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell]\!]}$ by notation a_i^ℓ. Further, the second-order derivative of the log-partition function with respect to parameters Θ_i^ℓ and $\Theta_j^{\ell'}$ is as follows.

$$\frac{\partial^2}{\partial \Theta_i^\ell \Theta_j^{\ell'}} \log Z(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{x} \sim \mathsf{p}}[a_i^\ell a_j^{\ell'}]$$

Consider a_i^ℓ as random variables distributed according to distributions defined by parameter $\boldsymbol{\Theta}$. Then the Hessian matrix of $\log Z(\boldsymbol{\Theta})$ is the co-variance matrix of random variables a_i^ℓ which is positive definite. Therefore the log-partition function $\log Z(\boldsymbol{\Theta})$ is a convex function in parameters $\boldsymbol{\Theta}$. □

Proposition 3.2. *The log-likelihood function $L(\boldsymbol{\Theta}; \mathcal{D})$ is concave if the weight functions are log-linear in their parameters.*

Proof. This result directly follows from the linearity of the first term and the convexity of $\log Z(\boldsymbol{\Theta})$ (Prop. 3.1). □

The concavity of the log-likelihood function allows the problem of computing the maximum likelihood to be formulated as a convex minimization problem and optimally solved with gradient descent methods. The partial derivative of the log-likelihood with respect to a parameter Θ_ℓ^k is:

$$\begin{aligned} \frac{\partial}{\partial \Theta_\ell^k} L(\boldsymbol{\Theta}) = & \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[[\![\boldsymbol{x} \models \ell]\!] \cdot f_\ell^k(\boldsymbol{x})] \\ & - \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{X}; \boldsymbol{\Theta})}[[\![\boldsymbol{x} \models \ell]\!] \cdot f_\ell^k(\boldsymbol{x})] \end{aligned} \tag{7}$$

Notice that the first expectation is a weighted count of the dataset $\mathcal{D}$ and thus can be computed in $\Theta(|\mathcal{D}|)$. The second expectation requires the computation of the partition function, i.e. a WMI problem, which is intractable in general, and therefore, approximation algorithms are needed to compute the gradient updates. Recently, Zeng et al. [2020a] characterized the tractable class of WMI problems in terms conditions on both their structure and in the form of the weight functions [Zeng et al., 2020b]. A family of weight functions satisfies *tractable weight conditions* (TWC) iff:

i) it is closed under product;
ii) it admits efficient computation of antiderivatives;
iii) it is closed under definite integration over each variables.

Some examples when the weight function families satisfy TWCs include the cases where the features functions are constants, linear functions or log-polynomial functions. Next we present the previous result on the tractability of WMI problems which is necessary to the tractability of the partial derivatives in MLE.

Proposition 3.3. *[Zeng et al., 2020a] Let $\mathcal{WMI}(\boldsymbol{\Omega}, \log(n), tr)$ be the class of WMI problems with primal graph $\mathcal{G}$ with diameter of size $\Theta(\log(n))$ and treewidth tr. When $\mathcal{WMI}(\boldsymbol{\Omega}, \log(n), tr)$ has parametric weight function family Ω satisfies the TWCs, $\mathcal{WMI}(\boldsymbol{\Omega}, \log(n), tr)$ is a tractable WMI class for inference if-and-only-if treewidth $tr = 1$.*

Theorem 3.4. *If the WMI problem $\mathsf{WMI}(\Delta, \mathcal{W})$ is in the tractable WMI problem class $\mathcal{WMI}(\boldsymbol{\Omega}, \log(n), 1)$ and the feature functions $f_\ell^k(\boldsymbol{x})$ are in function family that satisfies TWCs, then the partial derivative in Equation 7 can be tractably computed.*

Proof. Given that the first term in Equation 7 can be computed linearly in $|\mathcal{D}|$, it suffices to show that the second term is a ratio of two tractable WMI problems:

$$\mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{X}; \boldsymbol{\Theta})}[[\![\boldsymbol{x} \models \ell]\!] \cdot f_\ell^k(\boldsymbol{x})] = \frac{\mathsf{WMI}(\Delta \wedge \ell, \mathcal{W}')}{\mathsf{WMI}(\Delta, \mathcal{W})},$$

where for all ℓ^*, $w'_{\ell^*} \in \mathcal{W}'$, $w_{\ell^*} \in \mathcal{W}$, if $\ell^* \neq \ell$, weight function $w'_{\ell^*} = w_{\ell^*}$; otherwise, the weight function is defined as $w'_{\ell^*}(\boldsymbol{x}) = w_{\ell^*}(\boldsymbol{x}) f_\ell^k(\boldsymbol{x})$. The tractability of $\mathsf{WMI}(\Delta \wedge \ell, \mathcal{W}')$ follows from the tractability of $\mathsf{WMI}(\Delta, \mathcal{W})$, given that they have the same primal graph and that $w'_{\ell^*}(\boldsymbol{x})$ is the product of two functions that satisfies the TWC. Therefore, the partial derivative in Equation 7 can be efficiently computed. □

With the results above, we characterized which families of distributions that are amenable to WMI-based inference admit both tractable and exact MLE of their parameters. In the following, we present a novel reduction of marginal inference in HL-MRFs to WMI and generalize these results to this popular class of models.

4 MARGINAL INFERENCE FOR HL-MRFS VIA WMI

Hinge-loss Markov Random fields (HL-MRFs) [Bach et al., 2017] are a recently proposed statistical relational learning framework used to model highly structured data. This family of models allows tractable and exact MAP inference when weighted constraints are defined using Łukasiewicz logic, a fuzzy relaxation of Boolean logic. However, exact marginal inference for HL-MRFs is intractable in general [Embar et al., 2019]. Therefore, the weight learning via MLE for HL-MRFs approximates expectations with MAP states. We investigate the connection between HL-MRFs and WMI problems and propose a reduction from HL-MRFs to WMI models. This reduction is profound in two senses. Firstly, it makes the HL-MRFs amenable to the MLE-based parameter learning approach as proposed in Section 3. Secondly, it allows the characterization of tractability of marginal inference for HL-MRFs with the conditions reported in Prop. 3.3. To the best of our knowledge, *this is the first class of tractable HL-MRFs for exact marginal inference.*

4.1 HINGE-LOSS MARKOV RANDOM FIELDS

Before describing the reduction to WMI, we provide the main definitions in HL-MRFs. W.l.o.g, in the following we assume that all the random variables are unobserved.

Definition 4.1. *(potential) Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be a vector of n variables with joint domain $\boldsymbol{\mathcal{D}} = [0,1]^n$. Let $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)$ be a vector of m continuous* ***potentials*** *of the form*

$$\phi_j(\boldsymbol{x}) = (\max\{\ell_j(\boldsymbol{x}), 0\})^{p_j} \tag{8}$$

where ℓ_j is a linear function of $\boldsymbol{X}$ and $p_j \in \{1, 2\}$.

Typically, ℓ_j are interpreted as clauses in Łukasiewicz logic, with $\phi_j(\boldsymbol{x})$ representing the (possibly squared) distance to the satisfaction of the linear constraint $\ell_j \leq 0$.

Definition 4.2. *(constrained hinge-loss energy function) For $\boldsymbol{X} \in \boldsymbol{\mathcal{D}}$, given a vector of m non-negative free parameters, i.e., weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}^m$, a* ***constrained hinge-loss energy function*** *$f_{\boldsymbol{\alpha}}$ is defined as follows.*

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{m} \alpha_j \phi_j(\boldsymbol{x}). \tag{9}$$

This energy function quantifies the (weighted) satisfaction of the linear constraints above. These soft constraints are complemented with a set of hard constraints (i.e. with infinite weights) that are defined separately for convenience.

Definition 4.3. *(feasible set) Let $\boldsymbol{c} = (c_1, \cdots, c_r)$ be a vector of r linear constraint functions associated with index sets denoting equality constraints $\mathcal{E}$ and inequality constraints $\mathcal{I}$, which define the* ***feasible set***

$$\tilde{\boldsymbol{\mathcal{D}}} = \left\{ (\boldsymbol{x}) \in \boldsymbol{\mathcal{D}} \;\middle|\; \begin{array}{l} \forall k \in \mathcal{E}, c_k(\boldsymbol{x}) = 0 \\ \forall k \in \mathcal{I}, c_k(\boldsymbol{x}) \leq 0 \end{array} \right\}. \tag{10}$$

Then, the probability density function is defined over the feasible set and is inversely proportional to the constrained hinge-loss energy function.

Definition 4.4. *(HL-MRF) A* ***hinge-loss Markov random field*** *p over random variables $\boldsymbol{X}$ is a probability density defined as follows: if $\boldsymbol{x} \notin \tilde{\boldsymbol{\mathcal{D}}}$, then $p(\boldsymbol{x}) = 0$; if $\boldsymbol{x} \in \tilde{\boldsymbol{\mathcal{D}}}$, then*

$$p(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \exp\left(-f_{\boldsymbol{\alpha}}(\boldsymbol{x})\right) \tag{11}$$

where $Z(\boldsymbol{\alpha})$ is the partition function, i.e.,

$$Z(\boldsymbol{\alpha}) = \int_{\boldsymbol{x} | \boldsymbol{x} \in \tilde{\boldsymbol{\mathcal{D}}}} \exp\left(-f_{\boldsymbol{\alpha}}(\boldsymbol{x})\right) d\boldsymbol{x}. \tag{12}$$

The *templated HL-MRFs* can be similarly defined, where the potentials are defined by the templates denoted by $\boldsymbol{\mathcal{T}} = (t_1, \cdots, t_S)$ where each t_s is the set of indices of the potentials defined by the template. The templates are associated with weights $\boldsymbol{W} = (W_1, \cdots, W_S)$. Then the sum of potentials defined by a template t_s is $\boldsymbol{\Phi}_s(\boldsymbol{x}) = \sum_{i \in t_s} \phi_i(\boldsymbol{x})$. For each potential $\phi_i(\boldsymbol{x})$ with its index i in t_s, they are assigned weights to be the weight of the template, that is, $\alpha_i = W_s$. Together, the constrained hinge-loss energy function of a templated HL-MRF is as follows.

$$f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{\Phi}(\boldsymbol{x}) \tag{13}$$

where $\boldsymbol{\Phi}(\boldsymbol{x}) = (\boldsymbol{\Phi}_1(\boldsymbol{x}), \cdots, \boldsymbol{\Phi}_S(\boldsymbol{x}))$. The templated HL-MRFs are often used in the parameter learning when the weights are considered to be shared among potentials [Bach et al., 2017].

4.2 FROM HL-MRFS TO WMI

This section presents the reduction from HL-MRFs to WMI models that bridge the two frameworks. It allows the inference on HL-MRFs to be amenable to the tractability results on WMI problems as well as WMI solvers.

Theorem 4.5. *For any HL-MRF p over random variables $\boldsymbol{X}$, there exists a WMI problem $\mathsf{WMI}(\Delta, \mathcal{W})$ with per-literal weights whose WMI density denoted by p_Δ equals to the HL-MRF p.*

Proof. Given an HL-MRF p, construct an SMT($\mathcal{LRA}$) formula Δ in CNF form as follows.

$$\begin{aligned} \Delta = & \underbrace{\bigwedge_{i \in [n]} (0 \leq x_i) \wedge (x_i \leq 1)}_{\text{joint domain } \Delta_{\boldsymbol{\mathcal{D}}}} \\ & \underbrace{\bigwedge_{k \in \mathcal{E}} (c_k(\boldsymbol{x}) = 0) \bigwedge_{k \in \mathcal{I}} (c_k(\boldsymbol{x}) \leq 0)}_{\text{hard constraints } \Delta_{\tilde{\boldsymbol{\mathcal{D}}}}} \end{aligned} \tag{14}$$

Then construct a set of per-literal weights $\mathcal{W}$ with the set of literals $\mathcal{L} = \{\ell_j(\boldsymbol{x}) \geq 0\}_{j \in [m]}$ as follows. For brevity, here the per-literal weight is denoted by $\mathcal{W}_j$ with subscript being the index of the $\mathcal{LRA}$ atom ℓ_j in its associated literal $\ell_j(\boldsymbol{x}) \geq 0$.

$$\mathcal{W}_j(\boldsymbol{x}) = \exp(-\alpha_j \ell_j^{p_j}(\boldsymbol{x})) \tag{15}$$

With the above notations, the WMI density defined by Equation 4 is as follows.

$$p_\Delta(\boldsymbol{x}) = \frac{1}{Z_\Delta(\mathcal{W})} \prod_{j \in [m]} \mathcal{W}_j(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell_j(\boldsymbol{x}) \geq 0]\!]}, \quad \boldsymbol{x} \models \Delta. \tag{16}$$

with

$$Z_\Delta(\mathcal{W}) = \int_{\boldsymbol{x} \models \Delta} \prod_{j \in [m]} w_j(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell_j(\boldsymbol{x}) \geq 0]\!]} \, d\boldsymbol{x}. \tag{17}$$

Moreover, the HL-MRF P in Definition 4.4 has an equivalent formulation by using indicator function.

$$p(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\alpha})}[\![\boldsymbol{x} \in \boldsymbol{\mathcal{D}}]\!] \cdot [\![\boldsymbol{x} \in \tilde{\boldsymbol{\mathcal{D}}}]\!] \cdot \exp\left(-f_{\boldsymbol{\alpha}}(\boldsymbol{x})\right) \quad (18)$$

Next we show that the un-normalized HL-MRF p is proportional to the WMI density p_Δ. Notice that $[\![\boldsymbol{x} \in \boldsymbol{\mathcal{D}}]\!] = [\![\boldsymbol{x} \models \Delta_{\boldsymbol{\mathcal{D}}}]\!]$ and $[\![\boldsymbol{x} \in \tilde{\boldsymbol{\mathcal{D}}}]\!] = [\![\boldsymbol{x} \models \Delta_{\tilde{\boldsymbol{\mathcal{D}}}}]\!]$ hold. Moreover, since sub-formula $\Delta_{\boldsymbol{S}}$ is always SAT, it holds that

$$[\![\boldsymbol{x} \in \boldsymbol{\mathcal{D}}]\!] \cdot [\![\boldsymbol{x} \in \tilde{\boldsymbol{\mathcal{D}}}]\!] = [\![\boldsymbol{x} \models \Delta_{\boldsymbol{\mathcal{D}}}]\!] \cdot [\![\boldsymbol{x} \models \Delta_{\tilde{\boldsymbol{\mathcal{D}}}}]\!] = [\![\boldsymbol{x} \models \Delta]\!]. \quad (19)$$

Further since it holds that $\max\{\ell_j(\boldsymbol{x}), 0\} = [\![\ell_j(\boldsymbol{x}) \geq 0]\!] \cdot \ell_j(\boldsymbol{x})$, we could rewrite the potentials in HL-MRF using indicator functions.

$$\phi_j(\boldsymbol{x}) = [\![\boldsymbol{x} \models \ell_j(\boldsymbol{x}) \geq 0]\!] \cdot \ell_j^{p_j}(\boldsymbol{x}) \quad (20)$$

By Equations 19 and Equation 20, the following equality on un-normalized HL-MRF p holds for any $\boldsymbol{x} \models \Delta$.

$$\begin{aligned} Z(\boldsymbol{\alpha})p(\boldsymbol{x}) &= \prod_{j \in [m]} \exp(-\alpha_j \phi_j(\boldsymbol{x})) \\ &= \prod_{j \in [m]} \exp(-\alpha_j \ell_j^{p_j}(\boldsymbol{x}))^{[\![\boldsymbol{x} \models \ell_j(\boldsymbol{x}) \geq 0]\!]} \\ &= \prod_{j \in [m]} \mathcal{W}_j(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell_j(\boldsymbol{x}) \geq 0]\!]} = Z_\Delta(\mathcal{W}) p_\Delta(\boldsymbol{x}) \end{aligned}$$

Thus it holds that $p(\boldsymbol{x}) \propto p_\Delta(\boldsymbol{x})$. By the same arguments, it can be shown that HL-MRF p and WMI density p_Δ have the same partition functions, which finishes the proof for $p = p_\Delta$. □

What immediately follows the reduction is the tractability of marginal inference on HL-MRFs. When the potentials in the HL-MRF model is linear, the resulting WMI problem has its weight function to be in the exponetiated linear function family that satisfies the TWCs [Zeng et al., 2020a]. This gives the first non-trivial class of tractable HL-MRFs for exact marginal inference so far.

Corollary 4.6. *Given a HL-MRF p whose potentials are as defined in Definition 4.1 with $p_i = 1$, if the WMI problem reduced from p has its primal graph to have tree-width one and diameter of $\Theta(\log(n))$, then the marginal inference of the HL-MRF p is tractable.*

4.3 MLE FOR HL-MRFS

Besides the tractable marginal inference for HL-MRFs, what follows the reduction is that the that the parameter learning of WMI problems can be leveraged for the parameter learning of the HL-MRFs, and our analysis on the optimality and the tractability of MLE approach naturally generalize from WMI to HL-MRFs. To derive MLE-based parameter learning, we first compute the partial derivative of the log-likelihood for the templated HL-MRFs in Equation 13 as follows.

$$\begin{aligned} \frac{\partial}{\partial W_s} L(\boldsymbol{\Theta}) = &-\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[[\![\boldsymbol{x} \models \wedge_{i \in t_s} \ell_i]\!] \cdot \boldsymbol{\Phi}_s(\boldsymbol{x})] \\ &+ \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{X};\boldsymbol{\Theta})}\left[[\![\boldsymbol{x} \models \wedge_{i \in t_s} \ell_i]\!] \cdot \boldsymbol{\Phi}_s(\boldsymbol{x})\right] \end{aligned} \quad (21)$$

In previous work, since the expectation over the templated HL-MRFs is intractable in general, the MAP state is used to approximate the expectation [Bach et al., 2017]. However, our HL-MRFs-to-WMI reduction provides an alternative way to compute the expectation by using WMI solvers on the reduced WMI problem. Specifically, the expectation can be computed exactly and tractably when an exact WMI solver is applicable [Morettin et al., 2021]. Moreover, if the reduced WMI problem satisfies the assumptions in Theorem 3.4, the partial derivative for the templated HL-MRFs above can be computed exactly in tractable time. Below we show one of such cases.

Proposition 4.7. *Given a HL-MRF p with the same assumptions as in Corollary 4.6, then the partial derivative in Equation 21 can be tractably computed.*

5 RELATED WORK

Although most of the research output has focused on purely continuous or purely discrete distributions, in recent years probabilistic modelling for hybrid domains has received increasing attention.

Undirected hybrid models with potentials in the exponential family have been investigated for pairwise distributions [Lee and Hastie, 2015] and more general cases [Yang et al., 2014]. Density Estimation Trees (DETs) [Ram and Gray, 2011] address the problem of learning a distribution with axis-aligned piecewise constant components. Mixed Sum-Product Networks (MSPNs) [Molina et al., 2018, Vergari et al., 2019] learn mixtures of categorical and univariate piecewise polynomial distributions. In contrast with our setting, this line of research focuses on learning probabilistic relationships and the proposed learning and inference techniques disregard the algebraic and logical structures in the distributions.

The maximum-likelihood estimation of the parameters of piecewise distributions with logical and algebraic constraints, here investigated in the log-linear case, was initially addressed for piecewise constants by Belle et al. [2015]. Constraint learning in SMT, which is closely related to the problem of estimating the support of a structured hybrid distribution from data, was addressed in both supervised [Kolb et al., 2018] and unsupervised Morettin et al. [2020] settings. The latter work additionally propose LARIAT, a full pipeline for learning hybrid structured distributions from data by renormalizing DETs or MSPNs with the learned

constraints. This approach heavily relies on greedy procedures for learning both the structure and the parameters of the model, while we focus on a principled way of estimating the parameters of these distributions via MLE.

While HL-MRFs were initially developed to efficiently answer MAP queries, approximate marginal inference has been investigated by Embar et al. [2019]. Besides approximating MLE with MAP states, parameter learning in these models has been addressed using maximum pseudolikelihood, large-margin estimation [Bach et al., 2017] and, more recently, with Bayesian optimization techniques [Srinivasan et al., 2020].

6 CONCLUSION

In this work we investigated in a rigorous and principled manner the problem of estimating the parameters of hybrid models that account for logical and algebraic relationships among variables. We showed that MLE can be solved optimally for piecewise distributions that are log-linear in their parameters and, leveraging recent theoretical insights on inference in this setting, is also tractable when the structure of the model satisfy certain conditions.

We extended the above results to HL-MRFs by presenting a novel reduction of marginal inference in HL-MRFs to WMI, thus identifying a subclass of models that admit tractable and exact marginal inference and MLE-based parameter learning.

For future work, we aim at developing and evaluating practical algorithms for HL-MRFs based on the presented reduction. While we showed that log-linear weight functions admit optimal MLE, whether it is a necessary condition is still an open problem. In this work we focused on estimating the parameters in the weight function, but we plan to investigate the problem of estimating the coefficients in the weighted literals, i.e. finding the optimal piecewise decomposition of the joint probability.

References

Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V. Nori. Fairsquare: Probabilistic verification of program fairness. *Proc. ACM Program. Lang.*, (OOPSLA):80:1–80:30, 2017.

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18: 1–67, 2017.

Clark Barrett, Leonardo de Moura, Silvio Ranise, Aaron Stump, and Cesare Tinelli. The SMT-LIB initiative and the rise of SMT. In *Proceedings of the 6th international conference on Hardware and software: verification and testing*, pages 3–3. Springer-Verlag, 2010.

Vaishak Belle, Andrea Passerini, and Guy Van den Broeck. Probabilistic inference in hybrid domains by weighted model integration. In *Proceedings of IJCAI*, pages 2770–2776, 2015.

Varun Embar, Sriram Srinivasan, and Lise Getoor. Tractable marginal inference for hinge-loss markov random fields. In *Third ICML workshop on Tractable Probabilistic Modeling*, 2019.

Samuel Kolb, Stefano Teso, Andrea Passerini, and Luc De Raedt. Learning smt (lra) constraints using smt solvers. In *IJCAI*, volume 18, pages 2333–2340, 2018.

Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

Alejandro Molina, Antonio Vergari, Nicola Di Mauro, Sriraam Natarajan, Floriana Esposito, and Kristian Kersting. Mixed sum-product networks: A deep architecture for hybrid domains. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Paolo Morettin, Samuel Kolb, Stefano Teso, and Andrea Passerini. Learning weighted model integration distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5224–5231, 2020.

Paolo Morettin, Pedro Zuidberg Dos Martires, Samuel Kolb, and Andrea Passerini. Hybrid probabilistic inference with logical and algebraic constraints: a survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Parikshit Ram and Alexander G Gray. Density estimation trees. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–635, 2011.

Tian Sang, Paul Beame, and Henry A Kautz. Performing bayesian inference by weighted model counting. In *AAAI*, volume 5, pages 475–481, 2005.

Sriram Srinivasan, Golnoosh Farnadi, and Lise Getoor. Bowl: Bayesian optimization for weight learning in probabilistic soft logic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10267–10275, 2020.

Antonio Vergari, Alejandro Molina, Robert Peharz, Zoubin Ghahramani, Kristian Kersting, and Isabel Valera. Automatic bayesian density analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5207–5215, 2019.

Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2014.

Zhe Zeng and Guy Van den Broeck. Efficient search-based weighted model integration. *Proceedings of UAI*, 2019.

Zhe Zeng, Paolo Morettin, Fanqi Yan, Antonio Vergari, and Guy Van den Broeck. Probabilistic inference with algebraic constraints: Theoretical limits and practical approximations. *Advances in Neural Information Processing Systems*, 33, 2020a.

Zhe Zeng, Paolo Morettin, Fanqi Yan, Antonio Vergari, and Guy Van den Broeck. Scaling up hybrid probabilistic inference with logical and arithmetic constraints via message passing. In *Proceedings of the International Conference of Machine Learning (ICML)*, 2020b.

A PROOFS

Proof of Proposition 3.1. Denote the i-th parameter in weight function $w_{\ell'}$ as $\Theta_i^{\ell'}$. The first-order derivative of the log-partition function $\log Z(\mathbf{\Theta})$ with respect to a parameter $\Theta_i^{\ell'}$ can be computed as

$$\frac{\partial}{\partial \Theta_i^{\ell'}} \log Z(\mathbf{\Theta}) = \frac{1}{Z(\mathbf{\Theta})} \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} Z(\mathbf{\Theta})$$
$$= \frac{1}{Z(\mathbf{\Theta})} \cdot [\int_x \prod_{\ell \neq \ell'} w_\ell(x)^{[\![x \models \ell]\!]} \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} w_{\ell'}(x)^{[\![x \models \ell']\!]} \, dx]$$
$$= \frac{1}{Z(\mathbf{\Theta})} \cdot [\int_x \prod_{\ell} w_\ell(x)^{[\![x \models \ell]\!]} \cdot \frac{1}{w_{\ell'}(x)^{[\![x \models \ell']\!]}} \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} w_{\ell'}(x)^{[\![x \models \ell']\!]} \, dx]$$
$$= \frac{1}{Z(\mathbf{\Theta})} \cdot [\int_x \prod_{\ell} w_\ell(x)^{[\![x \models \ell]\!]} \cdot [\frac{\partial}{\partial \Theta_i^{\ell'}} \log w_{\ell'}(x)^{[\![x \models \ell']\!]}] \, dx]$$

Notice that the last equation follows from the fact that $\frac{1}{w_{\ell'}(x)^{[\![x \models \ell']\!]}} \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} w_{\ell'}(x)^{[\![x \models \ell']\!]} = \frac{\partial}{\partial \Theta_i^{\ell'}} \log w_{\ell'}(x)^{[\![x \models \ell']\!]}$. Given the WMI density of x in Equation 4, we have that

$$\frac{\partial}{\partial \Theta_i^{\ell'}} \log Z(\mathbf{\Theta}) = \mathbb{E}_{x \sim p}[\frac{\partial}{\partial \Theta_i^{\ell'}} \log w_{\ell'}(x)^{[\![x \models \ell']\!]}]$$

For brevity, denote $\frac{\partial}{\partial \Theta_i^{\ell}} \log w_\ell(\boldsymbol{x})^{[\![\boldsymbol{x} \models \ell]\!]}$ by notation a_i^ℓ. Further, the second-order derivative of the log-partition function with respect to parameters Θ_i^ℓ and $\Theta_j^{\ell'}$ is as follows.

$$\frac{\partial^2}{\partial \Theta_i^{\ell'} \partial \Theta_j^{\ell''}} \log Z(\mathbf{\Theta}) = \frac{\partial}{\partial \Theta_j^{\ell''}} \frac{\partial}{\partial \Theta_i^{\ell'}} \log Z(\mathbf{\Theta})$$
$$= \frac{\partial}{\partial \Theta_j^{\ell''}} \frac{1}{Z(\mathbf{\Theta})} \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} Z(\mathbf{\Theta}) + \frac{1}{Z(\mathbf{\Theta})} \cdot \frac{\partial^2}{\partial \Theta_i^{\ell'} \partial \Theta_j^{\ell''}} Z(\mathbf{\Theta})$$

where the first term can be rewritten as

$$\frac{\partial}{\partial \Theta_j^{\ell''}} \frac{1}{Z(\mathbf{\Theta})} \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} Z(\mathbf{\Theta})$$
$$= -\frac{1}{Z^2(\mathbf{\Theta})} \frac{\partial}{\partial \Theta_j^{\ell''}} Z(\mathbf{\Theta}) \cdot \frac{\partial}{\partial \Theta_i^{\ell'}} Z(\mathbf{\Theta})$$
$$= -\mathbb{E}_{\boldsymbol{x} \sim p}[a_j^{\ell''}] \cdot \mathbb{E}_{\boldsymbol{x} \sim p}[a_i^{\ell'}]$$

and the second term, given that the weight functions are log-linear in the parameters, can be rewritten as

$$\frac{1}{Z(\mathbf{\Theta})} \cdot \frac{\partial^2}{\partial \Theta_i^{\ell'} \partial \Theta_j^{\ell''}} Z(\mathbf{\Theta})$$
$$= \frac{1}{Z(\mathbf{\Theta})} \cdot \frac{\partial}{\partial \Theta_j^{\ell''}} [\frac{\partial}{\partial \Theta_i^{\ell'}} Z(\mathbf{\Theta})]$$
$$= \mathbb{E}_{\boldsymbol{x} \sim p}[a_j^{\ell''} a_i^{\ell'}]$$

We can view a_i^ℓ as random variables distributed according to distribution defined by parameters $\mathbf{\Theta}$. Then the Hessian matrix of $\log Z(\mathbf{\Theta})$ is the covariance matrix of random variables a_i^ℓ which is positive semidefinite. Therefore, the log-partition function $\log Z(\mathbf{\Theta})$ is a convex function in $\mathbf{\Theta}$, which finishes our proof. □